Demonstrate Var & Bias Change for Underfit Using Linear Regression

Interested in estimating μ.

True model: $Y_i^{(0)} = \beta_1 x_i + \mu + e_i$, $i=1,\dots,n$, $e_i \sim N(0,\sigma^2)$.

Data generated from true model: $\{(x_i, Y_i^{(0)})\}_{i=1}^n$

Smaller model: $Y_i = \mu + e_i \implies \hat{\mu}_{smaller} = \overline{Y_i^{(0)}}$
(Less complex)

$$\text{Bias}(\hat{\mu}_{smaller}) = E[\hat{\mu}_s] - \mu = E\left[\frac{1}{n}\sum_{i=1}^n Y_i^{(0)}\right] - \mu$$

$$= \frac{1}{n}\sum_{i=1}^n (\beta_1 x_i + \mu) - \mu = \frac{\beta_1}{n}\left(\sum x_i\right)$$

$$\neq 0 = \text{Bias}(\hat{\mu}_{correct})$$

Hence, $|\text{Bias}(\hat{\mu}_{smaller})| > |\text{Bias}(\hat{\mu}_{correct})|$.

$$\text{Var}(\hat{\mu}_{smaller}) = \text{Var}\left(\frac{1}{n}\sum Y_i^{(0)}\right) = \frac{\sigma^2}{n}$$

$$\begin{bmatrix} \hat{\mu}_c \\ \hat{\beta}_{1,c} \end{bmatrix} = \hat{\beta}_{correct} = (X^TX)^{-1}X^T y, \quad \text{Cov}(\hat{\beta}_{correct}) = \begin{bmatrix} \text{Var}(\hat{\mu}_{correct}) & \square \\ \square & \square \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\text{Cov}(\hat{\beta}_{correct}) = \text{cov}(Ay) = E[(Ay - EAy)(Ay - EAy)^T] = A\,\text{Cov}(\underset{\sim}{y})A^T$$

$$= (X^TX)^{-1}X^T \underbrace{\text{Cov}(\underset{\sim}{y})}_{\sigma^2 I} X(X^TX)^{-1} = \sigma^2 (X^TX)^{-1}$$

$$(X^T X)^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = n^{-1} \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum x_i^2 \end{pmatrix}^{-1}$$

$$= n^{-1} \begin{pmatrix} \boxed{\frac{1}{n}\sum x_i^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}^T \Big/ \left[\frac{1}{n}\sum x_i^2 - (\bar{x})^2\right]$$

Hence, $Var(\hat{\mu}_{correct}) = \sigma^2 \frac{1}{n}\sum x_i^2 \Big/ \left(\sum x_i^2 - n\bar{x}^2\right)$

$$= \frac{\sigma^2}{n} \cdot \frac{\sum x_i^2}{\sum x_i^2 - n\bar{x}^2} \geq \frac{\sigma^2}{n} = Var(\hat{\mu}_{smaller}).$$